

Weekly Report

胡万祺

一、 本周工作

本周开始正式地进行农业大数据项目的工作。我们想着首先从农科院那边给的样本数据入手，并分工进行不同数据的分析和研究：

- 胡万祺：shp 数据解析、绘制。
- 王叙萌、章天野：农业知识文档数据文本挖掘、文本预处理。
- 王艺：集群 nosql 数据库搭建和性能测试。
- 陈俊：集群 sql 数据库搭建和性能测试。

1. 查找网上农业大数据资源

见附件 1《数据来源》

2. Shp 数据

农科院给的 shp 数据表示的是二维平面，就是常见的地图，和一般的省界线图形类似；除此之外还有对应的 dbf 数据保存着一些信息。数据样本中的“耕地空间数据”和“土壤图”都是面状（polygon）的数据，这些数据本身就是作好的图，里面包含信息，而且带了坐标和投影的，一般的 GIS 软件打开就能查看的农科院处理这些 GIS 数据一般是基于 arcgis desktop 进行查看里面的信息，显示它的图；而要在 web 中显示用 shp 数据绘制的图则可用 nodejs 的 shp 模块和前端绘制库。我在网上下了世界地图的国界线和中国的省市县三级分界线做测试，最后的绘制结果如下：



图 1：世界地图国界线

其中图三柯城土壤图是根据农科院给的数据绘制的。我拿结果图去跟农科院邓老师验证了一下，其中：绘制结果的坐标体系和投影体系不太正确；图中的细长的区域是河流水系；土壤图不是代表某一田块的信息，而是一些不规则的土壤类型区域组成的。

另外还把 dbf 文件内的信息简单地放到每个区域上面，当鼠标移上去的时候显示。但是由于 dbf 的解析的时候中文字符乱码，所以先转成了 csv 文件，再进行解析。

之后的目标是把图 1——图 2——图 3 按照 zoom in/zoom out 的交互结合起来。



图 2：中国省市县三级分界线

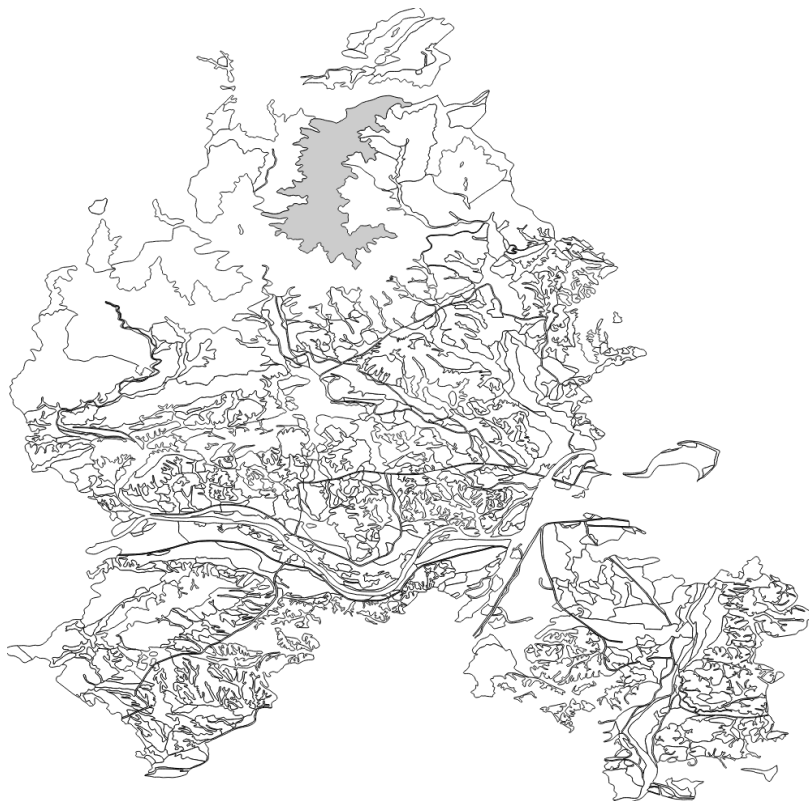


图 3：柯城土壤图

3. 集群数据库搭建和性能测试

● MongoDB

- 使用 `Mongodb + Nodejs` 对出租车二进制数据做了插入、创建索引测试。
- 使用异步操作效果仍不够理想，主要问题在于批量插入时有 1000 条数据限制。并且官方驱动 `bulk api` 效率不高，几乎和单条操作的循环效率差不多，使用这个 `api` 在乱序操作时为 45s/15 万条，顺序操作为 35s/15 万条。
- 使用 `insertMany` 操作时，同样受到每次 1000 条的限制，分片测试结果是 22s/15 万条插入总结：效率有待提升，可能需要换个驱动，或者用玄学手段突破 `node` 单线程限制。
- 至于索引还有空间使用，结果差不多，键名尽量简短可以缩小占用空间。

● PostgreSQL

在集群的一个节点机上部署了 `PostgreSQL` 数据库，并根据吴南师兄提供的一份出租车数据及 `java` 测试代码重建了该表，成功运行了测试代码，之后再将测试代码改写为 `nodejs` 的版本，运行成功。之后对两份代码都进行了大约 55000 条记录的插入测试，两者耗时均在 40s 左右，当然由于没有做大量的重复测试，这个数据应该不是很准确，但至少两者耗时的数量级应该是相同的。

另外直接跑 `java` 测试代码似乎也没有达到浩南师兄所描述的速度，猜测一方面由于测试的数据较少，每次批量插入的数据大小也变少了，另一方面则是运行环境差别较大。

4. 农业科普知识文档数据

上周给章天野和王叙萌布置的任务是：如何将一些已有的文本挖掘方面的理论基础及实际方法，应用到农业领域，即在我们已有的有关“病虫害”、“生产技术”等农业科普知识文档中，实现自动化处理文档、智能推荐等功能。她们非常认真地对这方面做了调研，她们的调研报告请见附件 2《调研报告-张天野&王叙萌》。